CS 7545: Machine Learning Theory **Fall 2021**

Lecture 3

Name: Tyler LaBonte **Professor:** Santosh S. Vempala

1 Singular Value Decomposition

Theorem 1.1: Singular Value Decomposition

Suppose $\mathbf{A} \in \mathbb{R}^{m \times d}$ and

$$\boldsymbol{v}_i \coloneqq \underset{\substack{\boldsymbol{v}:\boldsymbol{v}\perp\boldsymbol{v}_j, j\leq i \\ \|\boldsymbol{v}\|=1}}{\operatorname{argmax}} \|\mathbf{A}\boldsymbol{v}\|^2. \tag{1}$$

Then,

1. $V_k \coloneqq \operatorname{span}\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}$ is a subspace of dimension k that minimizes

$$\sum_{i=1}^{m} d(\boldsymbol{A}_i, \boldsymbol{v})^2. \tag{2}$$

   That is, V_k is a least squares subspace.

2. $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d$ are right singular vectors of $\mathbf{A}$ with corresponding left singular vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ and singular values $\sigma_1, \sigma_2, \ldots, \sigma_d$. Thus,

$$\mathbf{A} = \sum_{i=1}^{d} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\intercal. \tag{3}$$

Proof: The proof of 2 is an exercise (similar to $\boldsymbol{v}_1$ from last lecture). Here we prove 1 by induction on k. We showed $k = 1$ last time. Suppose V_{k-1} is a $(k-1)$-dimensional least squares subspace. Let V_k' be a k-dimensional least squares subspace. Let $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k$ be an orthonormal basis of V_k' such that $\boldsymbol{w}_k \perp V_{k-1}$. Then, V_k' maximizes

$$\sum_{i=1}^{k} \|\mathbf{A}\boldsymbol{w}_i\|^2 \leq \sum_{i=1}^{m} d(\boldsymbol{A}_i, V_{k-1})^2 + \|\mathbf{A}\boldsymbol{w}_k\|^2. \tag{4}$$

Since $\boldsymbol{w}_k$ was a candidate in the optimization of V_k,

$$\sum_{i=1}^{m} d(\boldsymbol{A}_i, V_{k-1})^2 + \|\mathbf{A}\boldsymbol{w}_k\|^2 \leq \sum_{i=1}^{m} d(\boldsymbol{A}_i, V_{k-1})^2 + \|\mathbf{A}\boldsymbol{v}_k\|^2 \tag{5}$$

$$= \sum_{i=1}^{m} d(\boldsymbol{A}_i, V_k)^2. \tag{6}$$

So V_k is a k-dimensional least squares subspace. ■

2 SVD and Mixtures of Gaussians

Theorem 2.1: SVD and Mixtures of Gaussians

Suppose F is a mixture of Gaussians with weights w_i and $F_i = \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 \mathbf{I}_d)$. Consider the algorithm which projects the sample to its top-k SVD subspace V_k, then clusters in $\mathbb{R}^k$ using distances (as in last lecture). Then,

1. V_k contains $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k\}$.
2. The algorithm suceeds with probability $1 - \delta$ if

$$\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\| \geq c\big(\log \frac{m}{\delta} \cdot k\big)^{1/4} \cdot \max(\sigma_i, \sigma_j). \tag{7}$$

Proof: 2 follows from 1 and last lecture. Here we prove 1. Suppose $k = 1$, then

$$\boldsymbol{v}_1 = \operatorname*{argmax}_{\|\boldsymbol{v}\|=1} \mathbb{E}_F[(\boldsymbol{x}^\intercal \boldsymbol{v})^2] \tag{8}$$

$$= \mathbb{E}\Big[((\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu}) \cdot \boldsymbol{v})^2\Big] \tag{9}$$

$$= \mathbb{E}\Big[((\boldsymbol{x} - \boldsymbol{\mu}) \cdot \boldsymbol{v})^2\Big] + (\boldsymbol{\mu} \cdot \boldsymbol{v})^2 \tag{10}$$

$$= \sigma^2 + (\boldsymbol{\mu} \cdot \boldsymbol{v})^2 \tag{11}$$

$$\implies \boldsymbol{v} = \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|}. \tag{12}$$

Because the Gaussians are spherical, the best k-dimensional subspace is any subspace containing $\boldsymbol{\mu}$. So for k Gaussians, we just contain all their means; this is optimal for each Gaussian individually, so it is optimal for their mixture. ■

Note that this strategy works for general Gaussians if the separation grows with the largest variance of the component Gaussians.

3 Linearly Independent Mixtures of Gaussians

What if we allow the Gaussians to overlap? Suppose we have a mixture F of k spherical Gaussians as above, and the only assumption we make is that $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ are linearly independent. We can't use clustering here, but perhaps we can estimate the parameters of the model. Here, second moments will not suffice.

We will need the property of *isotropy* to proceed. F is isotropic if $\mathbb{E}_F[\boldsymbol{x}] = \mathbf{0}$ and $\mathbb{E}_F[\boldsymbol{x}\boldsymbol{x}^\intercal] = \mathbf{I}$.

Fact 3.1: Isotropy

Any distribution with bounded second moments can be made isotropic by an affine transformation.

Proof: Suppose $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$ and $\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\intercal] = \mathbf{A} = \mathbf{B}\mathbf{B}^\intercal$. Let

$$\boldsymbol{y} = \mathbf{B}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \mathbf{A}^{-1/2}(\boldsymbol{x}-\boldsymbol{\mu}). \tag{13}$$

Then, $\mathbb{E}[\boldsymbol{y}] = \mathbf{0}$ and

$$\begin{aligned} \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\intercal] &= \mathbf{B}^{-1}\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\intercal](\mathbf{B}^{-1})^\intercal & (14)\\ &= \mathbf{B}^{-1}\mathbf{B}\mathbf{B}^\intercal(\mathbf{B}^{-1})^\intercal & (15)\\ &= \mathbf{I}. & (16) \end{aligned}$$

$\blacksquare$